AARHUS UNIVERSITY

Dear Editor,

Please find below a list of all relevant changes made to the manuscript during the revision process. To facilitate the review process, we have organized these changes according to the specific referee whose feedback prompted the revision.

Thank you for your continued consideration of our manuscript for publication in *Atmospheric Chemistry and Physics*.

Best regards,

Ivo Neefjes

AARHUS UNIVERSITY

List of Relevant Changes

Changes in Response to Referee 1

- **Kinetic Framework Clarification (L66 P3):** Explicitly defined the macroscopic cluster distribution dynamics framework used in atmospheric modeling (e.g., ACDC). Justified the assumption of a pressure-independent collision rate coefficient by noting the thermalization of large precursor clusters.
- **Methodological Scope (L80 P3):** Added explicit language emphasizing that this study is methodological, aimed at validating the long-range physical accuracy of machine learning interatomic potentials (MLIPs) prior to large-scale nucleation simulations.
- **GFN1-xTB Limitations (L38 P2):** Specified the types of errors inherent to GFN1-xTB for complex hydrogen-bonded systems, explicitly mentioning quantitative inaccuracies in binding energies and qualitative structural misidentifications.
- **Methodology Adaptability (L164 P6):** Added a note clarifying that while ωB97X-3c was used here, the workflow can be adapted to calculate atomic forces using any high-level quantum chemistry method on the GFN1-xTB structures.

Changes in Response to Referee 2

- **Topological Overlap Assumption (L157 P6):** Added a discussion detailing the implications of using GFN1-xTB dynamics to sample configurations for the ωB97X-3c potential. Explicitly noted that this assumes sufficient topological overlap and that missing configurations would require explicit augmentation.
- **Clarification of "Errors" (Table 3, Figure 2):** Updated captions to explicitly state that the reported MAE and RMSE values are "reproduction errors" relative to the specific level of theory the models were trained on.
- **PMF Evaluation Range (Table 4, L318 P12):** Adjusted the reporting of PMF RMSEs to strictly evaluate the region between the onset of the short-range repulsive wall and the long-range asymptotic zero-crossing. Excluded the highly repulsive, physically inaccessible regime (< 2.8 Å) to prevent sparse training data from skewing the reported accuracy.
- **Formatting and Typographical Fixes:**
  - Corrected "performance is consistently low" to "the errors are consistently low" (L286 P11).
  - Implemented a LaTeX float barrier to ensure Figure 2 and Table 3 appear in the correct chronological order.

Changes in Response to Professor Patrick Rinke (Referee 3)

- **Restructured Introduction (L29 P2):** Reorganized the introduction to introduce molecular dynamics (MD) and its specific challenges much earlier, improving logical flow.

- **PaiNN Charge Awareness (L107 P4):** Clarified that standard PaiNN is not explicitly charge-aware and handles ions by implicitly learning local structural effects. Noted that it treats molecules separated by more than its 10 Å cutoff as a disconnected, non-interacting graph.
- **AIMNet2 Architecture Updates (L115 P5, L126 P5):** Completely revised the AIMNet2 description to emphasize its use of invariant radial symmetry functions, generalized embeddings, and Neural Charge Equilibration (NQE) for explicit long-range Coulombic physics. Also removed an unsupported anecdotal claim regarding data efficiency comparisons with PaiNN.
- **Delta-Learning Citations and Bottlenecks (L131, L134 P5):** Added supporting citations (Ramakrishnan et al., 2015; Bogojeski et al., 2020) and clarified that Δ-learning efficiency is fundamentally bottlenecked by the need to evaluate the lower-level baseline at every time step.
- **PaiNN Spatial Cutoff Artifact (L291 P11):** Rewrote the paragraph explaining PaiNN's long-range failure. Clarified how the model erroneously maps long-range stabilization energy onto the slight internal geometric distortions of the isolated monomers due to its spatial cutoff.
- **Contextualizing Rate Coefficient Accuracy (L390 P18):** Added a paragraph explaining that because evaporation rates depend exponentially on binding free energies, an error factor of 1.5 in collision rate coefficients is highly acceptable for downstream atmospheric modeling. *(Note: This addition also addresses a similar comment made by Referee 2).*
- **Epoch Definitions (L247 P9):** Clarified that PaiNN uses the standard definition of an epoch (one full dataset pass), whereas AIMNet2 operationally defines an epoch as a user-specified number of steps, decoupling it from dataset size.
- **Hyperparameter Optimization (L264 P10):** Acknowledged that random grid search may not result in optimal hyperparameterd and referenced Bayesian optimization (Stuke et al., 2021) for future complex training tasks.
- **Clarifications and Visual Guidance:**
  - Spelled out "Intergovernmental Panel on Climate Change" (IPCC) (L20 P1).
  - Defined "gradient calculations" as the computation of atomic forces with respect to nuclear coordinates (L155 P6).
  - Added an early reference to PMF figures to help readers visualize the concepts of "well depth" and "shoulder" (L303 P12).
  - Corrected "great weight" to "greater weight" (L240 P9).

AARHUS UNIVERSITY

Dear Editor and Referee 1,

We sincerely thank referee 1 for evaluating our manuscript. However, we believe there is a fundamental misunderstanding of the manuscript's goal and methodology. In this cover letter, we outline the root of this disconnect, followed by detailed point-by-point responses to all specific comments below.

In the point-by-point responses, referee comments are given in ***bold italic***, while responses are given in roman (non-bold, non-italic). Excerpts from the revised manuscript to support our responses are highlighted in yellow. Line and page number are indicated by (L### P#).

We believe the referee's main criticisms stem from a fundamental misalignment regarding the manuscript's objective and its theoretical framework. First, this study is a methodological evaluation assessing how well machine learning models reproduce the long-range collision dynamics relevant to new particle formation. This evaluation is a mandatory prerequisite for accurately modeling large-scale nucleation simulations. Throughout the report, the referee critiques the study for evaluating systems below the critical cluster size, at a single temperature, and at an insufficiently high level of theory. However, these critiques apply to phenomenological kinetic studies rather than this methodological study. Because the underlying potential energy surface does not change with temperature, and the essential physics of monomer collisions must be validated before scaling to critical cluster sizes, exploring these additional avenues, while interesting, evaluates the manuscript against criteria outside its intended methodological scope.

Second, regarding the methodology, the referee approaches our work from the perspective of small-molecule gas-phase kinetics, where unstable excited complexes form and must be energetically evaluated for dissociation via a master equation. Atmospheric particle formation, on the other hand, standardly relies on a macroscopic cluster distribution dynamics formalism. This framework operates on the assumption that dissociation due to collisional excitation is negligible (Elm et al., 2020). This is physically motivated by the large number of internal degrees of freedom in the clusters studied here, which effectively accommodate excess collision energy until background gas thermalizes them. Therefore, the referee's disagreement lies fundamentally with the established theoretical framework of aerosol physics, rather than with our specific study. Given that macroscopic cluster dynamics simulations demonstrate good agreement with atmospheric experiments (e.g., Almeida et al., 2013), we do not believe this disagreement regarding standard field practices renders our study unfit for *Atmospheric Chemistry and Physics*.

This study represents a critical step toward utilizing machine learning interatomic potentials to run nucleation simulations at near-*ab initio* accuracy. Its focus on aerosol particle formation utilizing the research activity of machine learning aligns exactly with the Aims and Scope of *Atmospheric Chemistry and Physics*. While we have revised the manuscript to ensure its objectives and methodological framework is clear to future readers, the criticisms offered here do not accurately reflect the goals of this study or its situation within the broader atmospheric chemistry modeling community. As such, we respectfully request the Editor to carefully review our responses and consider our revised manuscript for publication in *Atmospheric Chemistry and Physics*.

Best regards,

Ivo Neefjes

AARHUS UNIVERSITY

Point-by-point response

Summary

Calculating the capture rate requires the long-range potential energy surface (PES) to be accurately defined. This paper uses molecular dynamics simulations (trajectory calculations) to explore the capture rate coefficient for three systems:

- ***$H_2SO_4 + H_2SO_4$; binding energy 66 kJ mol^{-1}***
- ***$H_2SO_4 + NH(CH_3)_2$; binding energy 71 kJ mol^{-1}***
- ***$H_2SO_4 + HSO_4^-$: binding energy 125 kJ mol^{-1}***

2 Machine learning architectures have been used to explore the PES for these systems based on data from molecular dynamics simulations at 2 levels of theory, semi-empirical and DFT.

Comparison is made between the Machine Learning methods and how things need to be improved to calculated accurate capture rates. All the capture rate coefficients are fast at 300 K.

We thank the referee for this summary of our study.

General criticism

This paper is mainly about the method of calculating the capture rate coefficient rather than the importance of the capture rate coefficients under consideration. If I wanted to calculate these capture rate coefficients, would it not be easier to do a high-level PES, much higher level than in this study, and then use capture rate theory to calculate the capture rate coefficient?

We thank the referee for their assessment and agree that this manuscript is primarily methodological. Its main objective is to investigate the extent to which machine learning algorithms can reproduce the attractive long-range interactions that govern sticking collisions and subsequent particle formation. We demonstrate that while all evaluated methods reproduce the free energy profile along the center-of-mass distance, only AIMNet2 successfully reproduces the collision rate coefficients due to its inclusion of long-range Coulombic and dispersion interactions.

While mapping a static, high-level potential energy surface (PES) might be feasible for a simple dimer collision, simulating the dynamics of spontaneous nucleation involves tens of flexible precursor molecules forming clusters of unpredictable sizes and configurations. Mapping a pre-defined PES for such a highly dimensional, multi-molecular system is computationally unfeasible. Molecular dynamics (MD) overcomes this bottleneck by sampling the PES on-the-fly, providing a cost-effective approach for unbiased cluster simulations.

Additionally, while methods like long-range transition state theory can provide collision rate coefficients at high levels of theory, they cannot account for explicit collision dynamics. Molecular

dynamics, in contrast, allows us to track the time-evolution of the system, capturing phenomena such as proton transfers, preferred orientations, anharmonic vibrations, and interconversions between low-lying free energy minima.

Furthermore, as demonstrated in this study, machine learning interatomic potentials approach the accuracy of their underlying quantum chemical training data at a fraction of the computational cost. While we utilized ωB97X-3c due to its proven performance in recent benchmarks (e.g., Neefjes et al., 2026), the methodology is highly adaptable. Models can be readily trained at even higher levels of theory by performing nuclear gradient calculations on structures sampled at the GFN1-xTB level. By validating the long-range accuracy of these models, this manuscript lays the necessary groundwork for large-scale nucleation simulations, where particle formation rates and collision rate coefficients are derived directly from the MD trajectories using methods such as the mean first-passage time (MFPT; Wedekind et al. (2007)) or Yasuoka-Matsumoto (Yasuoka and Matsumoto, 1998) approaches.

We have revised the manuscript to highlight its methodological nature and objective:

In this methodological study, we assess the ability of the PaiNN and AIMNet2 architectures to describe collisions governed by long-range interactions. We sampled training configurations using GFN1-xTB dynamics, subsequently computing energies and forces at both the GFN1-xTB and ωB97X-3c levels. Additionally, we employed delta-learning to upscale GFN1-xTB simulations with PaiNN corrections to the ωB97X-3c level of theory. Since sulfuric acid is a key contributor to particle formation (Sipilä et al., 2010), we studied the sulfuric acid dimer, the sulfuric acid–dimethylamine system (to investigate stabilizing proton transfers), and the sulfuric acid–bisulfate system (to examine strong ionic long-range contributions). Following hyperparameter tuning, we evaluated model performance by comparing electronic energy and force predictions against independent test sets. Furthermore, we calculated the potential of mean force (PMF) through umbrella sampling to compare against GFN1-xTB reference data. Finally, we derived collision rate coefficients from MD collision trajectory simulations to evaluate the long-range dynamics of the models and examine how they vary across different levels of theory. By validating these ML models in the context of atmospheric particle formation, this study establishes the necessary groundwork for large-scale MD simulations in this domain.

L80 P3

An even more important point is that these capture rate coefficients are not the values for atmospheric chemistry. The calculated capture rate coefficient is for infinite pressure and I'm very confident that the systems explored here are nowhere near the high-pressure limit. The value required for atmospheric chemistry requires a Master Equation calculation using the derived PES for the system. It is also likely, the adducts are too weakly bound for them to not re-dissociate back to reagents at 300 K. Therefore, the equilibrium coefficient for these systems might also be important. You have not provided the atmospherically relevant rate coefficient and the equilibrium coefficient for the system.

The calculated collision rate coefficients are specifically intended for use in macroscopic clus-

ter distribution models, such as those relying on birth-death equations (e.g., the Atmospheric Cluster Dynamics Code; McGrath et al. (2012)). In the birth-death framework, the collision process is explicitly separated from the evaporation process, operating under the assumption that nascent clusters live long enough to be thermalized by the background gas (Elm et al., 2020).

This assumption is physically justified by the size of the molecules likely involved in atmospheric new particle formation. The probability of fragmentation due to imperfect thermalization decreases rapidly as the number of vibrational modes ($3N - 6$, where N is the number of atoms) increases. Because the relatively large, strongly bound dimers studied here possess a large number of internal degrees of freedom to distribute the collision energy over. Consequently, the assumption of thermalization without dissociation at pressures encountered in the lower atmosphere (0.1 to 1 atm) is reasonable and represents standard practice in aerosol physics.

While small gas-phase radicals that form unstable complexes may dissociate before thermalization, thus requiring a master equation to calculate a pressure-dependent effective rate, this is not the kinetic regime of the aerosol precursors modeled here. Furthermore, evaporation rate coefficients and equilibrium constants for these systems can be, and frequently are, obtained via detailed balance using binding free energies from static quantum chemistry calculations. However, providing survival probabilities for these specific dimers is not the objective of this methodological study, which focuses strictly on evaluating whether machine learning models can accurately capture the long-range collision kinetics required for large-scale molecular dynamics simulations. We therefore respectfully maintain that we have provided the exact rate coefficients relevant to our stated objectives.

We have revised the manuscript to highlight the cluster distribution dynamics framework used and its assumption of a pressure-independent collision rate coefficient:

In cluster distribution dynamics models, such as the Atmospheric Cluster Dynamics Code (ACDC) (McGrath et al., 2012), cluster-forming collisions and cluster-removing evaporations are treated as independent processes, assuming that dissociation prior to thermalization from collisional excitation is negligible (Elm et al., 2020). This yields a pressure-independent collision rate coefficient, which represents the frequency of collisions per unit concentration.

L66 P3

The above two points raises the question of if ACP is the right journal for this paper. I do not think it is.

The referee's critique appears to stem from the framework of small-molecule fundamental gas kinetics, where master equation simulations are strictly required to track the dissociation of excited complexes. However, in the field of atmospheric new particle formation (NPF), this micro-kinetic framework is conventionally not applied, as large cluster precursors are physically understood to possess sufficient internal degrees of freedom to accommodate excitation energy.

By rejecting this assumption, the referee is inadvertently challenging the standard, widely accepted practice of modeling cluster distribution dynamics via macroscopic birth-death equations. Addressing the validity of foundational NPF kinetic frameworks falls far outside the scope of

this study, which aims specifically to evaluate the long-range physical accuracy of machine learning potentials, with the collision rate coefficient serving as the primary metric of success. This evaluation is a necessary preparatory step for large-scale, machine-learning-driven nucleation studies currently being developed within our group and the broader computational atmospheric chemistry community.

Atmospheric Chemistry and Physics has a history of publishing impactful research on both cluster distribution dynamics modeling (e.g., McGrath et al., 2012; Ortega et al., 2012) and the calculation of fundamental collision rate coefficients (e.g., Halonen et al., 2019; Neefjes et al., 2022; Yang et al., 2023). Because our work directly provides validated computational tools designed specifically to advance atmospheric nucleation modeling, we strongly believe it is suitable for publication in *Atmospheric Chemistry and Physics*.

Additionally, can you really explore the long-range potential with such a low level of theory?

Yes, because the physics of long-range molecular interactions are fundamentally different from short-range chemical bonding. At long distances, the potential energy surface (PES) is dominated by asymptotic electrostatic forces (such as dipole–dipole and ion–dipole interactions) and dispersion. These macroscopic molecular properties, specifically dipole moments and polarizabilities, converge very rapidly with the level of theory. Methods like ωB97X-3c and even semi-empirical GFN1-xTB are well-parameterized to reproduce these asymptotic electrostatic properties.

Furthermore, the adequacy of describing long-range potentials using these fundamental properties is strongly supported by decades of experimental evidence (e.g., Tsikritea et al., 2022). Experimental mass spectrometry and selected ion flow tube measurements of ion-molecule reaction rates consistently match the theoretical collision rate limits predicted by classical capture theories (such as the Su-Chesnavich model; Su and Chesnavich (1982)). These theoretical models rely entirely on the exact same basic long-range electrostatic parameters (polarizability and dipole moment) captured by our underlying computational methods.

Is it easy to calculate the capture rate coefficient over a range of temperatures? Only 300 K is presented.

Calculating collision rate coefficients across a range of temperatures is computationally straightforward, provided the underlying machine learning interatomic potential is properly trained for those conditions.

As temperature increases, thermal motion allows the system to access higher-energy conformational states and shorter intermolecular collision distances. If a machine learning (ML) model is trained exclusively on lower-temperature data, simulating higher temperatures risks pushing the model into extrapolative configurations where it may predict inaccurate forces. Therefore, exploring a wide temperature range requires ensuring the training set is adequately augmented with data sampled at or above the maximum target temperature.

However, conducting a temperature-dependent kinetic analysis falls outside the methodological

scope of this study. Our primary objective is to evaluate the fundamental ability of different ML architectures to capture long-range interactions, such as electrostatics and dispersion. The accuracy of the predicted potential energy surface (PES) at long distances is an intrinsic property of the ML architecture and its training and independent of the macroscopic simulation temperature. Because the underlying long-range PES does not change with temperature, benchmarking the models' physical accuracy at a single atmospherically relevant temperature (300 K) is sufficient to fulfill the objectives of this study.

Overall

For the reasons I outline above I do not think ACP is the right journal for this paper; these reactions are going to be pressure dependent and a long way from the capture rate coefficient at atmospheric pressure. To make this an ACP paper more work is required in order to calculate the atmospherically relevant rate coefficients, including the equilibrium coefficient.

As detailed in our previous responses, we respectfully maintain that the collisions of these aerosol precursors operate at or near the high-pressure limit at 1 atm, rendering the pressure-independent collision rate coefficient the relevant metric for atmospheric macroscopic cluster models.

Beyond the kinetic framework, this manuscript addresses a critical and highly specific bottleneck in atmospheric chemistry: the accurate simulation of new particle formation (NPF). Historically, large-scale molecular dynamics (MD) simulations of atmospheric nucleation have relied on classical force fields due to the prohibitive computational cost of *ab initio* MD. However, classical potentials fundamentally fail to model dynamic short-range quantum effects, such as the proton transfers that are essential for the thermodynamic stabilization of atmospheric acid–base clusters.

Recent advances in machine learning interatomic potentials (MLIPs) now allow us to run simulations that approach *ab initio* accuracy, explicitly capturing these atmospheric proton transfers, while maintaining accessible computational costs. However, before the atmospheric community can safely transition to machine-learning-driven nucleation simulations, it is essential to rigorously validate that these models do not sacrifice long-range physical accuracy (electrostatics and dispersion) to achieve short-range chemical accuracy. By evaluating this trade-off, this study represents an essential methodological step toward accurately simulating atmospheric nucleation systems.

Finally, regarding the manuscript's suitability, *Atmospheric Chemistry and Physics* explicitly lists machine learning as a research activity of interest within its Aims and Scope. Furthermore, our methodology directly provides the foundational kinetic inputs necessary for the macroscopic cluster distribution models (e.g., the Atmospheric Cluster Dynamics Code; (McGrath et al., 2012)) that have been extensively developed and published within *Atmospheric Chemistry and Physics* (e.g., Tuovinen et al., 2022; Shen et al., 2024). Because this study directly equips the atmospheric modeling community with validated computational tools to advance NPF research, we strongly believe it is suitable for publication in *Atmospheric Chemistry and Physics*.

AARHUS UNIVERSITY

Line-by-line

Are the collisions you are considering in this study important for NPF? I think it is later collisions that are rate determining, not the first step collisions considered in the present system.

The collision systems in this study were selected based on their fundamental importance to atmospheric new particle formation (NPF). Sulfuric acid is widely considered an essential driver of NPF due to its low volatility and its capacity to form strongly bound acid–base clusters. Among atmospheric amines, dimethylamine is recognized as the strongest candidate for such clustering, representing an optimal balance between atmospheric concentration, high basicity, and minimal steric hindrance. Furthermore, the bisulfate ion is a highly relevant component for modeling ion-induced particle formation pathways.

We acknowledge that, according to classical nucleation theory, the overall particle formation rate is governed by a critical cluster size, the threshold at which collisions begin to outweigh evaporation. Cluster distribution dynamic simulations (Olenius et al., 2013) and experiments (Kürten et al., 2014), however, show that atmospheric particle-forming systems, such as sulfuric acid–dimethylamine, do not have a critical cluster size but are collision controlled. Hence, the first collision forming the dimer is essentially the critical cluster.

Furthermore, reaching the critical cluster size is not the objective of this methodological study. Our aim is to evaluate, develop, and validate the computational tools necessary to accurately model the dynamics of nucleation using machine learning interatomic potentials. Because the entire macroscopic nucleation sequence is built upon a series of discrete collision events, accurately and efficiently describing the foundational monomer–monomer interactions is a mandatory first step. By benchmarking the machine learning models' ability to capture the long-range physics of these initial collisions, we are validating the tools required to eventually simulate the full pathway up to and beyond the critical cluster size.

The capture rate coefficient is appropriate at the high-pressure limit. A lot of associations—including the systems in this paper—are going to be pressure dependent at atmospheric pressure and will be much smaller than the capture limit. A Master Equation calculation is required.

As mentioned above, the framework of macroscopic cluster distribution dynamics assumes that the particle-forming precursors investigated in this study experience negligible dissociation due to collisional excitation. This is physically grounded in the structural complexity of the clusters. For instance, a sulfuric acid–dimethylamine dimer is relatively strongly bound and possesses 45 internal degrees of freedom across which the excess collision energy can be shared.

Consequently, the assumption of negligible dissociation (high-pressure limit) is not an approximation unique to our manuscript, but rather a foundational principle underpinning the established kinetic theory driving cluster dynamics modeling throughout the atmospheric sciences community.

What sort of errors. Do you just mean rate coefficients?

For the mentioned errors, we refer to a broader range of inaccuracies inherent to the semi-empirical GFN1-xTB method when applied to complex hydrogen-bonded systems, rather than errors in the collision rate coefficients themselves. Prominent examples of these limitations include both quantitative thermodynamic inaccuracies and qualitative structural misidentifications.

While GFN1-xTB offers exceptional computational efficiency, it can exhibit significant quantitative errors in binding energies. For instance, in our previous benchmarking study on hydrated clusters (Neefjes et al., 2026), GFN1-xTB yielded a mean electronic energy error of approximately 5 kcal/mol for various $(\text{acid/base})_{0-2}(\text{water})_{0-5}$ systems when compared to a highly accurate DLPNO-CCSD(T0)/aug-cc-pVTZ reference. For larger aerosol precursors, such as a cluster containing 6 sulfuric acid, 6 ammonia, and 10 water molecules, these quantitative electronic energy errors can exceed 40 kcal/mol against the same reference.

Furthermore, as an example of qualitative errors, GFN1-xTB frequently predicts a different lowest-energy conformational isomer (global minimum) compared to density functional theory methods, such as ωB97X-3c (Kubečka et al., 2019). Accurate evaporation rates depend fundamentally on correct thermodynamic stabilities and structural conformations.

We revised the manuscript to explicitly mention the type of errors we are alluding to:

While semi-empirical quantum chemistry methods such as GFN1-xTB (Grimme et al., 2017) can model bond-breaking, they can exhibit significant errors for complex hydrogen-bonded systems, including quantitative inaccuracies in binding energies (Neefjes et al., 2026) and qualitative misidentifications of lowest-energy cluster configurations (Kubečka et al., 2019).

L38 P2

Can MD simulations be run long enough to be relevant for more atmospheric reactions? Are you stuck at pico-second processes or shorter? For instance, adding H_2O to sulfuric acid dimer is important and the question is how many waters are added (together with further reactions) can MD tackle this problem, or is the timescale to long for MD? If you have an accurate PES from a much higher-level QM calculation, can this PES be used to calculate an accurate rate coefficient via capture rate theory?

We thank the referee for raising these practical considerations. With recent advances in machine learning interatomic potentials (MLIPs), the accessible timescale for *ab initio*-quality molecular dynamics (MD) has been significantly extended. While actual *ab initio* MD is often restricted to the picoseconds regime, MLIP-driven MD can simulate systems of tens to hundreds of molecules over durations of several to tens of nanoseconds.

Regarding the specific example of sequential water addition to a sulfuric acid dimer: due to the low concentration of particle-forming precursors compared to background gas in the actual atmosphere, simulating a realistic macroscopic slice of the atmosphere to observe these slow, sequential collisions in real-time is indeed impossible. However, MD studies routinely

overcome this timescale limitation using well-established methodologies. Direct nucleation simulations employ highly supersaturated conditions to observe clustering events within nanosecond timescales. The resulting particle formation rates can then be rigorously extrapolated down to atmospherically relevant concentrations.

Beyond direct nucleation, targeted MD setups, such as the isolated collision trajectory simulations utilized in this study, allow us to extract precise collision rate coefficients and capture complex dynamical effects that are often neglected in static theoretical frameworks.

Finally, regarding the use of a high-level quantum mechanical potential energy surface (QM PES) with capture theory: as discussed in our previous response, while a static QM PES might be feasible for a simple monomer–monomer collision, mapping the multi-dimensional QM PES required for sequential additions (such as a dimer interacting with water molecules) quickly becomes computationally infeasible. MD bypasses this dimensionality explosion by evaluating the PES on-the-fly exactly where the dynamics visits.

The systems considered can be calculated at a much higher level via standard ab initio techniques. Would it not be better to do this as well, in order to really assign the accuracy of the present calculations?

The aim of this study is not to benchmark the accuracy of the employed quantum chemistry methods. In several benchmarks, ωB97X-3c has been found to provide accurate energetics and configurations compared to higher level ab initio techniques (Jensen and Elm, 2024; Neefjes et al., 2026). Our concern is if the chosen quantum chemistry method is accurately reproduced by the machine learning model.

We have, however, revised the manuscript to highlight that the same methodology can be used to train machine learning models at higher levels of theory:

We note that while ωB97X-3c is used in this study, atomic forces can be calculated using any quantum chemistry method on the GFN1-xTB structures to obtain a training set at that level of theory.

L164 P6

This is not absolute chemical accuracy. What are you comparing against to state errors below 0.1 kcal mol^{-1}?

Defining "chemical accuracy" as an energy error of less than 1 kcal/mol is standard nomenclature within the computational chemistry community. This was likely established by John Pople. As stated in his Nobel lecture (Pople, 1999): "As the model becomes quantitative, the target should be that data is reproduced and predicted within experimental accuracy. For energies, such as heats of formation or ionization potentials, a global accuracy of 1 kcal/mole would be appropriate."

Beyond its comparison with experimental uncertainty, this 1 kcal/mol threshold also has implications for kinetic and thermodynamic modeling. Free energies dictate equilibrium constants

and evaporation rate coefficients via exponential relationships. At 300 K, an energy error of approximately 1.37 kcal/mol translates to a full order-of-magnitude error in these exponential quantities. Therefore, an energy error less than 1 kcal/mol ensures that rate and equilibrium errors are restricted to less than a factor of ten.

Regarding the specific values reported in our manuscript, the stated errors (below 0.1 kcal/mol) do not represent absolute accuracy compared to experimental "truth". Rather, they quantify how well the machine learning interatomic potential reproduces the potential energy surface of the specific quantum chemical level of theory (ωB97X-3c) it was trained on. An error below 0.1 kcal/mol confirms that the machine learning architecture introduces negligible additional error, effectively preserving the fundamental chemical accuracy of the underlying training data.

The highest observed RMSE is 0.20 kcal mol^{-1} for AIMNet2 applied to the H_2SO_4–HSO_4^- system, which is five times lower than the standard threshold for chemical accuracy (1 kcal mol^{-1}). In other words, the reproduction error introduced by the machine learning models is negligible compared to generally accepted error margins in computational chemistry.

L334 P14

References

Almeida, J., Schobesberger, S., Kürten, A., Ortega, I. K., Kupiainen-Määttä, O., Praplan, A. P., Adamov, A., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Dommen, J., Donahue, N. M., Downard, A., Dunne, E., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Guida, R., Hakala, J., Hansel, A., Heinritzi, M., Henschel, H., Jokinen, T., Junninen, H., Kajos, M., Kangasluoma, J., Keskinen, H., Kupc, A., Kurtén, T., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Leiminger, M., Leppä, J., Loukonen, V., Makhmutov, V., Mathot, S., McGrath, M. J., Nieminen, T., Olenius, T., Onnela, A., Petäjä, T., Riccobono, F., Riipinen, I., Rissanen, M., Rondo, L., Ruuskanen, T., Santos, F. D., Sarnela, N., Schallhart, S., Schnitzhofer, R., Seinfeld, J. H., Simon, M., Sipilä, M., Stozhkov, Y., Stratmann, F., Tomé, A., Tröstl, J., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Virtanen, A., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Williamson, C., Wimmer, D., Ye, P., Yli-Juuti, T., Carslaw, K. S., Kulmala, M., Curtius, J., Baltensperger, U., Worsnop, D. R., Vehkamäki, H., and Kirkby, J.: Molecular understanding of sulphuric acid–amine particle nucleation in the atmosphere, Nature, 502, 359–363, https://doi.org/10.1038/nature12663, 2013.

Elm, J., Kubečka, J., Besel, V., Jääskeläinen, M. J., Halonen, R., Kurtén, T., and Vehkamäki, H.: Modeling the formation and growth of atmospheric molecular clusters: A review, J. Aerosol Sci., 149, 105 621, https://doi.org/10.1016/j.jaerosci.2020.105621, 2020.

Halonen, R., Zapadinsky, E., Kurtén, T., Vehkamäki, H., and Reischl, B.: Rate enhancement in collisions of sulfuric acid molecules due to long-range intermolecular forces, Atmos. Chem. Phys., 19, 13 355–13 366, https://doi.org/10.5194/acp-19-13355-2019, 2019.

Jensen, A. B. and Elm, J.: Massive Assessment of the Geometries of Atmospheric Molecular Clusters, J. Chem. Theory Comput., 20, 8549–8558, https://doi.org/10.1021/acs.jctc.4c01046, 2024.

Kubečka, J., Besel, V., Kurtén, T., Myllys, N., and Vehkamäki, H.: Configurational Sampling of Noncovalent (Atmospheric) Molecular Clusters: Sulfuric Acid and Guanidine, J. Phys. Chem. A, 123, 6022–6033, https://doi.org/10.1021/acs.jpca.9b03853, 2019.

Kürten, A., Jokinen, T., Simon, M., Sipilä, M., Sarnela, N., Junninen, H., Adamov, A., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., Dommen, J., Donahue, N. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Hakala, J., Hansel, A., Heinritzi, M., Hutterli, M., Kangasluoma, J., Kirkby, J., Laaksonen, A., Lehtipalo, K., Leiminger, M., Makhmutov, V., Mathot, S., Onnela, A., Petäjä, T., Praplan, A. P., Riccobono, F., Rissanen, M. P., Rondo, L., Schobesberger, S., Seinfeld, J. H., Steiner, G., Tomé, A., Tröstl, J., Winkler, P. M., Williamson, C., Wimmer, D., Ye, P., Baltensperger, U., Carslaw, K. S., Kulmala, M., Worsnop, D. R., and Curtius, J.: Neutral molecular cluster formation of sulfuric acid–dimethylamine observed in real time under atmospheric conditions, Proc. Natl. Acad. Sci. U.S.A., 111, 15 019–15 024, https://doi.org/10.1073/pnas.1404853111, 2014.

McGrath, M. J., Olenius, T., Ortega, I. K., Loukonen, V., Paasonen, P., Kurtén, T., Kulmala, M., and Vehkamäki, H.: Atmospheric Cluster Dynamics Code: a flexible method for solution of the birth-death equations, Atmos. Chem. Phys., 12, 2345–2355, https://doi.org/10.5194/acp-12-2345-2012, 2012.

Neefjes, I., Halonen, R., Vehkamäki, H., and Reischl, B.: Modeling approaches for atmospheric ion–dipole collisions: all-atom trajectory simulations and central field methods, Atmos. Chem. Phys., 22, 11 155–11 172, https://doi.org/10.5194/acp-22-11155-2022, 2022.

Neefjes, I., Knattrup, Y., Wu, H., Trolle, G. B., Elm, J., and Kubečka, J.: Thermodynamic benchmarking of hydrated atmospheric clusters in early particle formation, Aerosol Research, 4, 1–22, https://doi.org/10.5194/ar-4-1-2026, 2026.

Olenius, T., Kupiainen-Määttä, O., Ortega, I. K., Kurtén, T., and Vehkamäki, H.: Free energy barrier in the growth of sulfuric acid–ammonia and sulfuric acid–dimethylamine clusters, J. Chem. Phys., 139, 084 312, https://doi.org/10.1063/1.4819024, 2013.

Ortega, I. K., Kupiainen, O., Kurtén, T., Olenius, T., Wilkman, O., McGrath, M. J., Loukonen, V., and Vehkamäki, H.: From quantum chemical formation free energies to evaporation rates, Atmos. Chem. Phys., 12, 225–235, https://doi.org/10.5194/acp-12-225-2012, 2012.

Pople, J. A.: Nobel Lecture: Quantum chemical models, Rev. Mod. Phys., 71, 1267–1274, https://doi.org/10.1103/RevModPhys.71.1267, 1999.

Shen, J., Zhao, B., Wang, S., Ning, A., Li, Y., Cai, R., Gao, D., Chu, B., Gao, Y., Shrivastava, M., Jiang, J., Zhang, X., and He, H.: Cluster-dynamics-based parameterization for sulfuric acid–dimethylamine nucleation: comparison and selection through box and three-dimensional modeling, Atmos. Chem. Phys., 24, 10 261–10 278, https://doi.org/10.5194/acp-24-10261-2024, 2024.

Su, T. and Chesnavich, W. J.: Parametrization of the ion–polar molecule collision rate constant by trajectory calculations, J. Chem. Phys., 76, 5183–5185, https://doi.org/10.1063/1.442828, 1982.

Tsikritea, A., Diprose, J. A., Softley, T. P., and Heazlewood, B. R.: Capture theory models:

An overview of their development, experimental verification, and applications to ion–molecule reactions, J. Chem. Phys., 157, 060 901, https://doi.org/10.1063/5.0098552, 2022.

Tuovinen, S., Cai, R., Kerminen, V.-M., Jiang, J., Yan, C., Kulmala, M., and Kontkanen, J.: Survival probabilities of atmospheric particles: comparison based on theory, cluster population simulations, and observations in Beijing, Atmos. Chem. Phys., 22, 15 071–15 091, https://doi.org/10.5194/acp-22-15071-2022, 2022.

Wedekind, J., Strey, R., and Reguera, D.: New method to analyze simulations of activated processes, J. Chem. Phys., 126, 134 103, https://doi.org/10.1063/1.2713401, 2007.

Yang, H., Neefjes, I., Tikkanen, V., Kubečka, J., Kurtén, T., Vehkamäki, H., and Reischl, B.: Collision-sticking rates of acid–base clusters in the gas phase determined from atomistic simulation and a novel analytical interacting hard-sphere model, Atmos. Chem. Phys., 23, 5993–6009, https://doi.org/10.5194/acp-23-5993-2023, 2023.

Yasuoka, K. and Matsumoto, M.: Molecular dynamics of homogeneous nucleation in the vapor phase. I. Lennard-Jones fluid, J. Chem. Phys., 109, 8451–8462, https://doi.org/10.1063/1.477509, 1998.

AARHUS UNIVERSITY

Dear Editor and Referee 2,

We sincerely thank referee 2 for their clear, constructive, and valuable feedback. Their insightful comments helped us to recognize the specific areas where our manuscript could be strengthened and clarified. We have taken all of their constructive critiques to heart and have revised the manuscript accordingly.

In the following, we provide point-by-point responses to all their comments. Referee comments are given in ***bold italic***, while our responses are given in roman (non-bold, non-italic). Excerpts from the revised manuscript to support our responses are highlighted using yellow highlight. The line and page number to which a response refers is indicated by (L### P#).

We believe that the incorporated revisions and our detailed responses address the referee's feedback and strengthen the manuscript. We respectfully submit this revised version for your consideration for publication in *Atmospheric Chemistry and Physics*.

We look forward to hearing from you at your earliest convenience and thank you for considering our manuscript for publication.

Best regards,

Ivo Neefjes

AARHUS UNIVERSITY

1 Point-by-point responses

Neefjes et al. evaluate machine-learned interatomic potentials for molecular dynamics simulations of collisions between atmospherically relevant molecules. The study compares PaiNN, AIMNet2, and Δ-learning approaches trained on GFN1-xTB and ωB97X3c data. The authors show that models relying purely on local atomic environments can struggle to capture long-range interactions relevant for collision dynamics, whereas AIMNet2 performs well due to its explicit long-range treatment. While the importance of long-range interactions in MLIPs is not a new observation, the manuscript provides a useful and well-executed demonstration in the context of atmospheric chemistry. I recommend publication after the following revisions.

We thank the referee for their positive assessment and accurate summary of our study. We agree that the limitations of purely local machine learning interatomic potentials are well established within the broader computational chemistry community. However, as the referee notes, the specific consequences of these limitations on atmospheric particle formation modeling had not been previously investigated. We view this work as the necessary validation before these models can be deployed for large-scale nucleation simulations, demonstrating that they generally reproduce the underlying level of theory with low errors, while specifically confirming their ability to accurately capture the long-range forces governing collisions.

Page 5, line 114. "While it may not match the data efficiency or accuracy of PaiNN for geometry-sensitive properties..." Please provide a citation or quantitative evidence supporting this claim.

The referee is entirely justified in questioning this statement. Upon reflection, this claim was based on anecdotal observations during our group's work with these models, rather than a rigorous quantitative benchmark. Because we have not formally compared the relative data efficiencies of PaiNN and AIMNet2, and are unaware of published literature that provides this specific comparison, we agree that the statement lacks the necessary supporting evidence. Consequently, we have removed this sentence from the revised manuscript:

Designed for generalizability, AIMNet2 natively supports systems with different charge states and spin multiplicities. By explicitly accounting for these varying electronic states and long-range interactions, the model is well-suited for a wide range of chemically complex systems.

L126 P5

Page 4. Similarly, the discussion of PaiNN emphasizes its equivariant representation of vectorial properties. Does AIMNet2 include a comparable treatment of directional features, or does it rely on something different? Clarifying this would help readers compare the architectures.

We thank the referee for this question. It is indeed important to contrast PaiNN's use of equivariant directional features with the approach taken by AIMNet2. Unlike PaiNN, AIMNet2 does not enforce explicit equivariant representations for vectorial properties. Instead, it relies on invariant radial symmetry functions, capturing directional effects implicitly through its atom-centered environment representations and iterative message passing. We have now clarified this architectural distinction in the AIMNet2 method section:

AIMNet2 is the second generation of the Atoms-In-Molecules Neural Network developed by Anstine et al. (2025). Using a message-passing architecture, the model iteratively refines invariant representations of local atomic environments, defined by radial symmetry functions, to build complex "atom-in-molecule" (AIM) embeddings. While directional dependencies are not enforced through explicit equivariance, AIMNet2 captures these effects implicitly through its atom-centered representations and iterative message passing. A key feature of AIMNet2 is its generalized embedding strategy, which avoids element-specific subnetworks and allows the model to flexibly represent highly diverse chemical compositions.

L115 P5

Page 5, line 143. "Instead, it was assumed that the GFN1-xTB PES sufficiently overlaps with the relevant regions of the ωB97X-3c PES." I understand that MD with ωB97X-3c would be expensive. However, the manuscript would be improved by adding the implications of this decision.

We agree with the referee that the implications of this approach warrants proper discussion in the manuscript.

This dataset generation approach assumes that the topology of the GFN1-xTB PES qualitatively resembles that of the higher-level target method. Under this condition, minor structural discrepancies, such as slight differences in the bond lengths of local minima, are not problematic. The higher-level nuclear gradient calculations provide a net force directing the geometry toward the true ωB97X-3c minimum. Furthermore, while GFN1-xTB might incorrectly rank the relative energies of specific configurations, the ωB97X-3c calculations will correctly reorder them, provided all relevant local minima are sufficiently sampled during the semi-empirical dynamics.

The main limitation of this approach occurs if important regions of the ωB97X-3c configurational space are entirely missing from the GFN1-xTB PES. If the semi-empirical dynamics fails to visit specific critical conformations, the machine learning model will never be trained on them. In such cases, the training dataset would require explicit augmentation by sampling those missing configurations directly at the target level of theory.

For the relatively simple collision systems discussed here, this will most likely not be an issue, but for more complex clusters of flexible molecules, it is important to validate this assumption.

We have revised the manuscript to explicitly discuss these implications.

Instead, it was assumed that the GFN1-xTB PES sufficiently overlaps with the relevant regions of the ωB97X-3c PES. This assumption holds as long as the GFN1-xTB PES has

the same topological features as the ωB97X-3c PES in the relevant regions. Small structural discrepancies are corrected, as the higher-level nuclear gradients provide a net force directing the geometry toward the true ωB97X-3c minimum. The assumption, however, breaks down if important regions of the ωB97X-3c PES are entirely missing from the GFN1-xTB PES. While this is unlikely for the relatively simple collision systems studied here, more complex clusters may require the dataset to be augmented with unvisited structures.

L157 P6

In Figure 2 and Table 3, please clarify what the errors are relative to in the captions.

This is a fair point. We agree that it is easy for a reader to become confused about what the errors signify when the reference is not explicitly stated. The values shown in Figure 2 and Table 3 represent reproduction errors, i.e., the error of a machine learning model, trained on a specific level of theory, relative to calculations performed directly at that level of theory. To make this unambiguous, we have revised the captions as follows:

Table 3. Mean absolute errors (MAE) of the machine learning models relative to the level of theory they were trained on for electronic energies (E_{el}) and component-wise forces (F) across the three studied systems: H_2SO_4–H_2SO_4, H_2SO_4–$NH(CH_3)_2$, and H_2SO_4–HSO_4^-. Results are reported for AIMNet2, PaiNN, and Δ-PaiNN trained on GFN1-xTB and ωB97X-3c training data. Units: E_{el} in kcal mol^{-1} and F in kcal mol^{-1} Å^{-1}.

Figure 2. Electronic energy reproduction errors for machine learning models trained on ωB97X-3c data relative to the ωB97X-3c level of theory, shown as a function of the center-of-mass (COM) distance across the three studied systems: H_2SO_4–H_2SO_4 (a), H_2SO_4–$NH(CH_3)_2$ (b), and H_2SO_4–HSO_4^- (c). Results are shown for AIMNet2, PaiNN, and Δ-PaiNN.

I believe that Figure 2 isn't introduced in the text until after Table 3 is introduced. The authors should adjust the ordering of their manuscript.

We thank the referee for pointing this out. The inverted placement in the compiled manuscript was an artifact of LaTeX maintaining independent floating queues for figures and tables, which inadvertently allowed the table to bypass the earlier figure despite being placed correctly in our source code. We have implemented a float barrier in the revised manuscript to ensure they now appear in the correct order in the current single-column format. Should the manuscript be accepted, we will communicate this intended order to the typesetter during the proofing stage of the two-column layout.

Page 10, line 259. "For the H_2SO_4-H_2SO_4 system, performance is consistently low across the entire coordinate." This implies that performance is bad, when I think you mean to say that the errors are low across the entire coordinate.

This is indeed badly phrased. We have changed "performance is" to "the errors are":

For the H_2SO_4-H_2SO_4 system, the errors are consistently low across the entire coordinate.

L286 P11

Page 13, Table 4. It is my understanding that the Δ-PaiNN model here has been trained on GFN1-xTB as the baseline and target reference. I agree with the authors note on page 10 that this should yield an essentially-zero correction. Why then is the full RMSE of PaiNN 0.053 kcal/mol and the full RMSE of Δ-PaiNN 3.4 kcal/mol? I would expect these to be closer. If this difference is due to the "inherent uncertainty associated with finite umbrella sampling" (page 12), then should this have been properly mitigated or evaluated separately? Does this actually justify such a large error?

We thank the referee for raising this issue. It is indeed surprising that the Δ-PaiNN method does not reproduce the GFN1-xTB curve within a smaller margin of error, and an RMSE of 3.4 kcal/mol cannot be ascribed to uncertainties in the umbrella sampling simulations alone. To investigate this, we plotted the short-distance repulsive wall of the potential of mean force (PMF) for H_2SO_4–H_2SO_4 (Fig. R1). At energies above ∼30 kcal/mol, the PaiNN model perfectly follows the GFN1-xTB reference, while AIMNet2 and Δ-PaiNN erroneously bend toward a slower increase in energy. We have rigorously verified our data to ensure this is not an artifact of mislabeling the PaiNN and Δ-PaiNN models.

The energy deviation occurs at center-of-mass (COM) distances below 2.8 Å. Out of our 20,000-structure dataset, only 63 structures fall below this distance, and only 22 are below 2.5 Å. Although we sampled windows down to a COM distance of 2.0 Å, the large repulsive forces in this region counteract the bias potential. Consequently, the molecules rarely adopt configurations with COM distances below 2.8 Å, leaving the training set overwhelmingly biased toward larger distances. This bias is intentional, as our focus is on the system's behavior near the minimum and at long ranges. Accurately modeling the highly repulsive region would require a data generation scheme that explicitly targets this region.

To demonstrate that this region of the potential energy surface is physically negligible for our purposes, we evaluated the maximum kinetic energy available in our simulations for scaling the repulsive wall. We performed collision trajectory simulations with initial relative velocities up to 800 m/s, which corresponds to a kinetic energy of approximately 3.8 kcal/mol for the H_2SO_4–H_2SO_4 system. Even if this entire kinetic energy were converted to potential energy to climb the repulsive wall, the system would not reach the energy threshold where the curves begin to diverge.

Furthermore, we can calculate the standard-state (Helmholtz) binding free energy, $\Delta F^{\ominus}$, using the Δ-PaiNN curve as is versus a curve where the tail below 2.8 Å is corrected to perfectly follow the GFN1-xTB reference:

$$\Delta F^{\ominus} = -k_{\mathrm{B}}T \ln \frac{4\pi}{V^{\ominus}} \int_0^{r_{\max}} r^2 \exp\left(-\beta w(r)\right) \mathrm{d}r \,, \tag{1}$$

where k_B is the Boltzmann constant, T is the temperature, $V^\ominus$ is the standard volume, r is the COM distance, $\beta = 1/k_B T$, and $w(r)$ is the PMF. Integrating up to $r_{max} = 10$ Å yields -5.6154008 kcal/mol for the uncorrected Δ-PaiNN curve and -5.6154012 kcal/mol for the corrected curve. Thus, the deviation between GFN1-xTB and Δ-PaiNN in the highly repulsive region is entirely negligible for both equilibrium and collision properties.

The question remains as to why Δ-PaiNN deviates in the repulsive region rather than applying the expected near-zero correction to the GFN1-xTB baseline. Because of the extreme sparsity of training data at low COM distances, the model lacks the constraints necessary to uniquely identify the physically correct potential energy surface. Consequently, Δ-PaiNN overfits to an erroneous path in this unconstrained region, inappropriately adjusting the baseline GFN1-xTB results. While this artifact could be resolved by generating additional training data in the repulsive regime, we have refrained from retraining the models since the dynamics of this inaccessible region are outside the scope of our study.

Given the referee's valid criticism of the full RMSE, we propose removing the distinction between "full" and "shoulder" RMSE in the manuscript. Instead, we will only report the RMSE between the two points where the PMF crosses zero (i.e., in the repulsive and long-range regions). This modification does not alter our conclusions. Rather, it streamlines Table 4 and the accompanying text, ensuring that we only evaluate and discuss the regions of the potential energy surface for which the models were rigorously trained and which are physically accessible.

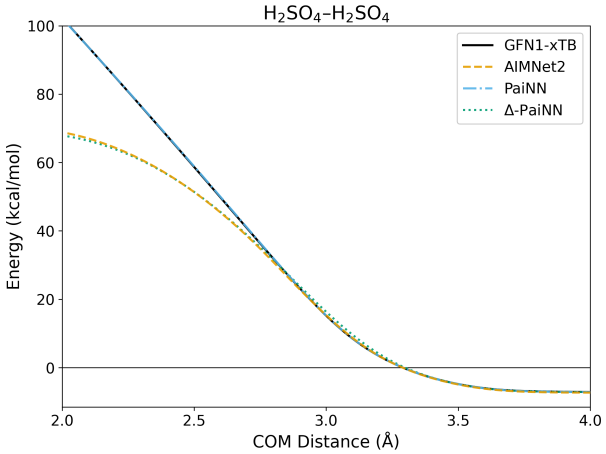

Figure R1: The repulsive region of the potential of mean force along the center-of-mass (COM) distance obtained from umbrella sampling for the H_2SO_4–H_2SO_4 collision system. Results are shown for GFN1-xTB and the AIMNet2, PaiNN, and Δ-PaiNN machine learning models trained on GFN1-xTB.

The revised table:

Table 1: Root-mean-square errors (RMSE) in kcal mol^{-1} for the potentials of mean force (PMFs) predicted by the machine learning models relative to the GFN1-xTB reference. The RMSE is evaluated between the point where the PMF drops below zero in the short-range repulsive region and the first point it returns to zero in the long-range non-interacting region.

System	Method	Evaluation Range (Å)	RMSE (kcal mol^{-1})
H_2SO_4–H_2SO_4	AIMNet2	3.28–11.06	0.15
	PaiNN		0.058
	Δ-PaiNN		0.042
H_2SO_4–$NH(CH_3)_2$	AIMNet2	2.72–8.78	0.046
	PaiNN		0.17
	Δ-PaiNN		—
H_2SO_4–HSO_4^-	AIMNet2	2.90–15.00	0.20
	PaiNN		0.090
	Δ-PaiNN		—

The revised accompanying text:

Figure 3 shows the PMFs for all three systems calculated using AIMNet2 and PaiNN (trained on GFN1-xTB), compared against the GFN1-xTB reference. For the H_2SO_4–H_2SO_4 system, the Δ-PaiNN model trained on GFN1-xTB was again included as a sanity check. While this model should theoretically reproduce the GFN1-xTB reference PMF, small deviations are nonetheless expected due to the inherent uncertainty associated with finite umbrella sampling. Additionally, because the machine learning models lack training data in the highly repulsive, physically inaccessible regimes at short distances, much larger deviations can occur in these regions (see Fig. S3 in the Supporting Information).

Table 4 lists the root-mean-square errors (RMSEs) of the predicted PMFs relative to the reference. We report the RMSE only between the point where the PMF drops below zero in the short-range repulsive region and the first point it returns to zero in the long-range non-interacting region. At shorter distances, the steep energies of the repulsive wall lead to sparse training data coverage, which can result in localized high errors. However, because these configurations are physically inaccessible at atmospheric temperatures, excluding this region ensures the reported RMSE reflects the model's performance in the region relevant to the clustering dynamics. Conversely, we exclude the asymptotic long-range tail because the collision partners are essentially non-interacting here. Including an extensive non-interacting region, which is well-sampled and exhibits minimal energy variation, would disproportionately lower the average error, masking the model's performance in the interaction region. While the PMF should theoretically approach zero asymptotically, sampling noise causes the long-range zero-crossing to occur at finite distances (< 20 Å) for the systems studied here.

L318 P12

Page 17, Table 5. The text regarding this figure would benefit from identifying what magnitude of error in the collision rate coefficients is considered acceptable for atmospheric modeling applications.

We thank the referee for this comment. It is indeed important to contextualize the obtained values. The accuracy of particle formation rates obtained from cluster distribution dynamics simulations depends on the accuracy of the collision rate coefficients and the cluster binding free energies used to calculate the evaporation rate coefficients. Because evaporation rate coefficients depend exponentially on the cluster binding free energy, an error of just 1 kcal/mol already results in a discrepancy of a factor of ∼5. As such, errors in binding free energies typically outweigh errors in collision rate coefficients. To keep errors in particle formation rates below an order of magnitude in the worst-case scenario where the evaporation and collision errors compound in the same direction, and assuming a binding free energy error of around chemical accuracy (1 kcal/mol), we consider an error of up to a factor of 1.5 in the collision rate coefficient to be acceptable.

We have added a paragraph explaining this context to the manuscript.

To contextualize these results, it is important to note that the accuracy of particle formation rates in cluster distribution dynamics simulations depends on both the collision and evaporation rate coefficients. Because evaporation rates depend exponentially on binding free energies, these errors typically outweigh errors in collision rate coefficients. An error of just 1 $\mathrm{kcal\,mol^{-1}}$ in binding free energies introduces a factor of ∼5 uncertainty in the evaporation rate. As such, we consider an error of a factor of 1.5 in the collision rate coefficients acceptable. In the worst-case scenario where collision and evaporation rate coefficient errors compound in the same direction, a factor of 1.5 collision rate coefficient error would still result in an overall uncertainty in the particle formation rates of less than an order of magnitude.

Evaluated against this threshold, PaiNN yields notably lower rate coefficients for the charged H_2SO_4–HSO_4^- system. For both GFN1-xTB and ωB97X-3c training data, the model underestimates the rates by nearly 50% (roughly a factor of 2) relative to the GFN1-xTB reference. As discussed in Sec. 3.5, this substantial deviation stems from the model's inability to detect collisions beyond its 10 Å cutoff, effectively neglecting the significant contribution of the long-range tail.

Conversely, for the neutral systems (H_2SO_4–H_2SO_4 and H_2SO_4–$NH(CH_3)_2$), the ML models trained on GFN1-xTB data exhibit excellent agreement with the reference calculations. All three architectures reproduce the GFN1-xTB reference rate coefficients closely, with the largest deviation observed for AIMNet2 applied to the H_2SO_4–$NH(CH_3)_2$ system (∼10% discrepancy).

L390 P18

AARHUS UNIVERSITY

Dear Editor and Professor Patrick Rinke,

We sincerely thank the referee for their comprehensive feedback. Their comments have helped us improve the manuscript, particularly by enhancing the robustness of our machine learning methodology description and making the text more accessible to a broader audience. We have carefully considered all of their constructive suggestions and adjusted the manuscript accordingly.

Below, we provide a point-by-point response to each comment. The referee's comments are given in ***bold italics***, while our responses are provided in standard roman text. Excerpts from the revised manuscript are indicated with a yellow highlight to support our responses. The line and page number to which a response refers is denoted by (L### P#).

We believe these incorporated revisions and our detailed responses address the referee's feedback and strengthen the manuscript. We respectfully submit this revised version for your consideration for publication in *Atmospheric Chemistry and Physics*.

We look forward to hearing from you at your earliest convenience and thank you for considering our manuscript for publication.

Best regards,

Ivo Neefjes

AARHUS UNIVERSITY

1 Point-by-point responses

The manuscript evaluates machine-learning interatomic potentials for simulating gas-phase collisions and early clustering among key atmospheric precursors (sulfuric acid with itself, dimethylamine, and bisulfate) linking short-range reactive accuracy to long-range interaction fidelity that controls capture and collision rates. Using umbrella sampling and umbrella integration to reconstruct potentials of mean force (PMF) with explicit subtraction of the radial entropic term, the authors derive collision probabilities and rate coefficients from molecular dynamics and analyze how model architecture affects both kinetics and thermodynamics. The simulations capture reactive events such as proton transfer during acid-base encounters, while the conclusions stress that gains in short-range accuracy must not compromise the long-range forces that govern initial approach and rate enhancement; accordingly, the authors urge validation beyond scalar error metrics (MAE/RMSE) and propose deploying models with explicit long-range interactions where collision kinetics are targeted.

For Atmospheric Chemistry and Physics (ACP) readers, the study offers a practical framework to obtain physically faithful collision-sticking rates and PMFs for nucleation-relevant systems and provides actionable guidance on model selection: local, short-range-accurate networks for thermodynamic sampling and global or long-range–aware models for encounter kinetics, consistent with recent ACP advances on collision-sticking analysis. The data, trained models, and scripts are openly available via the Atmospheric Cluster Database, facilitating reproducibility and uptake in nucleation and early growth modelling. I therefore recommend publication once the more detailed comments below have been addressed.

We thank the referee for their positive assessment and for taking the time to provide such a detailed and accurate summary of our study.

"IPCC assessment report" - It would be good to spell out IPCC

We agree with the referee that while the abbreviation IPCC is well known in the atmospheric sciences community, we hope this paper will also be read by the machine learning and general quantum chemistry communities. As such, we have clarified this abbreviation:

According to the latest Intergovernmental Panel on Climate Change (IPCC) assessment report,...

L20 P1

"These calculations explicitly account for long-range interactions and provide a fully atomistic description. Furthermore, the resulting trajectories offer insight into the molecular-level dynamics governing collisions and the formation of stable clusters." There is a logical disconnect in the introduction. After this sentence all

seems fine, but I guess that is, because the paragraph leading up to this sentence did not specify how the MD was calculated. Maybe one can allude the reader already to the forthcoming problems otherwise the introduction meanders along for quite a while until the readers reaches the actual problem definition.

We thank the referee for their careful review and for pointing out this logical disconnect. We agree that our initial effort to cover the relevant MD literature inadvertently obscured the overarching logical thread of the introduction.

To improve the narrative flow, we have restructured this section to introduce molecular dynamics, and the specific challenges of using it to study particle formation, much earlier, specifically in the third paragraph. While the core content remains largely unchanged, we believe this revised structure establishes the study's primary focus much sooner and directly addresses the referee's concern:

To capture the dynamic nature of these initial steps, researchers increasingly rely on atomistic molecular dynamics (MD) simulations. These simulations provide a fully atomistic description and offer insight into the molecular-level dynamics governing collisions and the formation of stable clusters. However, accurately capturing the necessary physics at a reasonable computational cost remains challenging. The quality of MD simulations is in large part determined by the level of theory at which the interaction potential between the nuclei in the system is obtained. An ideal interatomic potential for particle formation MD must satisfy three competing requirements: it must accurately capture long-range attractive forces to model how molecules initially approach; it must describe short-range quantum effects, such as chemical reactions, to model cluster stabilization; and it must be computationally efficient enough to sample a statistically significant number of events.

These effects include chemical reactions like proton transfers, which play a critical role in stabilizing atmospheric clusters. While semi-empirical quantum chemistry methods such as GFN1-xTB (Grimme et al., 2017) can model bond-breaking, they can exhibit significant errors for complex hydrogen-bonded systems, including quantitative inaccuracies in binding energies (Neefjes et al., 2026) and qualitative misidentifications of lowest-energy cluster configurations (Kubečka et al., 2019). Higher levels of theory, such as the DFT composite method ωB97X-3c (Müller et al., 2023), offer the necessary short- and long-range accuracy, but their computational cost makes even short MD simulations of small systems prohibitively expensive. Thus, neither classical nor conventional ab initio methods fully satisfy the requirements for large-scale cluster formation MD.

Recently, several machine learning (ML) architectures have been developed to construct accurate interatomic potentials for molecular systems. These machine learning interatomic potentials (MLIPs) offer a potential solution to this tradeoff, promising to reproduce the accuracy of high-level quantum theory at a reasonable computational cost. For instance, the polarizable atom interaction neural network (PaiNN) is an equivariant message-passing neural network capable of accelerating MD simulations while maintaining accuracy comparable to its reference training data (Schütt et al., 2021; Kubečka et al., 2024). Similarly, the second-generation atoms-in-molecules neural network (AIMNet2) has demonstrated high predictive accuracy across a wide range of molecular systems with remarkable efficiency, enabling

simulations of systems containing up to 10^5 atoms (Anstine et al., 2025).

However, MLIPs often rely on a local atomic environment approximation, in which the model encodes the environment around each atom up to a user-defined cutoff radius. This approximation improves transferability and computational efficiency but inherently limits the model to short-range interactions. The PaiNN model addresses this through a message-passing framework, where atoms exchange information with their neighbors via message and update blocks. Through multiple iterations, the effective interaction range grows, allowing atoms to indirectly access information from beyond the immediate cutoff. However, if all atoms in one subsystem (e.g., a molecule) lie beyond the cutoff radius of another, the interaction graph becomes disconnected. Consequently, no messages are exchanged, and the model treats the subsystems as non-interacting. AIMNet2 mitigates this by supplementing message passing with explicit long-range contributions. It predicts partial charges to model analytical Coulomb interactions and adds dispersion effects via the D3(BJ) correction scheme (Grimme et al., 2010, 2011).

The potential of MLIPs for atmospheric modeling has already been demonstrated by several recent studies that simulated the evolution of systems containing tens of particle-forming molecules to observe cluster formation dynamics (Jiang et al., 2022, 2023; Liu and Jiang, 2025). As the field increasingly adopts these methods for large-scale simulations, it is important to evaluate how well different model architectures capture long-range interactions alongside the necessary short-range accuracy and computational efficiency.

A rigorous metric for evaluating this long-range capability is the canonical collision rate coefficient. In cluster distribution dynamics models, such as the Atmospheric Cluster Dynamics Code (ACDC) (McGrath et al., 2012), cluster-forming collisions and cluster-removing evaporations are treated as independent processes, assuming that dissociation prior to thermalization from collisional excitation is negligible (Elm et al., 2020). This yields a pressure-independent collision rate coefficient, which represents the frequency of collisions per unit concentration. Traditionally, this coefficient is calculated using kinetic gas theory. In this framework, colliding partners are approximated as hard spheres, and intermolecular interactions are neglected entirely. While analytical approaches like the central field model can account for long-range forces, they require interaction parameters that are significantly more difficult to determine than standard hard-sphere radii (Neefjes et al., 2025).

Atomistic MD collision trajectory simulations in the free molecular regime provide a powerful alternative to calculate these coefficients directly from the underlying physical interactions (Halonen et al., 2019; Neefjes et al., 2022; Yang et al., 2023; Knattrup et al., 2025; Tikkanen et al., 2025). As demonstrated by Halonen et al. (2019), explicitly capturing long-range interactions via MD using the classical OPLS-AA force field resulted in an enhancement factor of 2.7 relative to kinetic gas theory for sulfuric acid dimerization. Accurately reproducing these enhanced collision rates serves as a robust metric for evaluating the long-range behavior of MLIPs in atmospheric applications.

In this methodological study, we assess the ability of the PaiNN and AIMNet2 architectures to describe collisions governed by long-range interactions. We sampled training configurations using GFN1-xTB dynamics, subsequently computing energies and forces at both the GFN1-xTB and ωB97X-3c levels. Additionally, we employed delta-learning to upscale GFN1-xTB

simulations with PaiNN corrections to the ωB97X-3c level of theory. Since sulfuric acid is a key contributor to particle formation (Sipilä et al., 2010), we studied the sulfuric acid dimer, the sulfuric acid–dimethylamine system (to investigate stabilizing proton transfers), and the sulfuric acid–bisulfate system (to examine strong ionic long-range contributions). Following hyperparameter tuning, we evaluated model performance by comparing electronic energy and force predictions against independent test sets. Furthermore, we calculated the potential of mean force (PMF) through umbrella sampling to compare against GFN1-xTB reference data. Finally, we derived collision rate coefficients from MD collision trajectory simulations to evaluate the long-range dynamics of the models and examine how they vary across different levels of theory. By validating these ML models in the context of atmospheric particle formation, this study establishes the necessary groundwork for large-scale MD simulations in this domain.

L29 P2

PaiNN section: Is PaiNN actually charge aware? How does PaiNN handle the anionic system in the study?

We thank the referee for this insightful question, which directly relates to the limitations observed in our results. The standard PaiNN architecture lacks explicit, physics-based charge awareness. While the total molecular charge can be provided to PaiNN as an additional descriptor during training, the model treats it merely as a learned feature. Unlike AIMNet2, which uses charge to compute explicit long-range Coulombic interactions, PaiNN does not contain the built-in physics to propagate electrostatic effects beyond their defined cutoff radius.

PaiNN handles the charged system by implicitly learning the local effects of the charge directly from the training data. Because the reference semi-empirical and DFT calculations inherently capture the stronger interactions, polarization, and altered local geometries of the charged state, PaiNN successfully reproduces the complex short-range potential energy surface of the charged system. However, because it lacks explicit charge awareness and long-range Coulombic terms, it is blind to electrostatic interactions that extend beyond its 10 Å cutoff.

We have updated the methodology section to discuss this charge-unawareness in more detail:

The standard PaiNN architecture is not explicitly charge-aware and lacks long-range electrostatic or dispersion corrections. Instead, the model handles charged systems by implicitly learning the local, short-range effects of the charge directly from the energies and forces provided in the training data. However, a limitation of this purely local approach in the context of gas-phase collisions is that interactions cannot be transmitted between atoms separated by more than the cutoff distance without intermediate atoms to mediate the message passing. Increasing the cutoff can capture these long-range effects, but at the expense of higher computational cost and potentially reduced accuracy, as the model must learn to generalize over a significantly larger spatial domain.

L107 P4

AIMNet2 section: "The model combines local atomic environments with learned "atom-in-molecule" (AIM) embeddings" — I don't understand why these two aspects are singled out for AIMNet2. All message passing graph neural networks do this. Also PaiNN uses a sophisticated input representation of the local atomic environments and then builds up atomic embeddings during the training.

AIMNet2 section: "These embeddings, available for 14 elements..." — Are the authors speaking of an AIMNet2 foundation or a pre-trained model? Otherwise it shouldn't matter what the elements are, because a pristine AIMNet2 architecture could be trained on the data at hand and then contain the elements present in that dataset.

The referee is completely correct in questioning these statements. Coming from a computational chemistry background, our original text did not properly describe the machine learning architecture. We inadvertently conflated the fundamental mechanics of standard message-passing networks, as well as the specific parameterization of the pre-trained models, with the architectural specifics of AIMNet2. As the referee rightly points out, iteratively building environment-aware embeddings is a feature of all message-passing neural network potentials (including PaiNN), and the 14-element parameterization refers to the pre-trained weights rather than a limitation of the architecture. To correct this, we have thoroughly revised the AIMNet2 methodology section. We removed the confusing statements and instead focused on the true architectural distinctions of AIMNet2 that are relevant to our study: its generalized embedding strategy, native charge awareness, and reliance on radial symmetry functions:

AIMNet2 is the second generation of the Atoms-In-Molecules Neural Network developed by Anstine et al. (2025). Using a message-passing architecture, the model iteratively refines invariant representations of local atomic environments, defined by radial symmetry functions, to build complex "atom-in-molecule" (AIM) embeddings. A key feature of AIMNet2 is its generalized embedding strategy, which avoids element-specific subnetworks and allows the model to flexibly represent highly diverse chemical compositions.

Beyond its local representations, AIMNet2 explicitly incorporates electronic and long-range physical effects. The architecture is charge-aware, using the total molecular charge as an input parameter to dynamically infer atom-centered partial charges during the message-passing phase. These partial charges are iteratively updated through a neural charge equilibration (NQE) scheme. The total potential energy is then calculated as the sum of the local configurational energy, explicit Coulombic electrostatic interactions derived from the learned partial charges, and a D3(BJ) dispersion correction.

Designed for generalizability, AIMNet2 natively supports systems with different charge states and spin multiplicities. By explicitly accounting for these varying electronic states and long-range interactions, the model is well-suited for a wide range of chemically complex systems.

L115 P5

AIMNet2 section: "While it may not match the data efficiency or accuracy of PaiNN for geometry-sensitive properties..." — Citations would be warranted to

back up this statement, unless it refers to the conclusions of this study, in which case this should be stated.

The referee is entirely justified in questioning this statement. Upon reflection, this claim was based on anecdotal observations during our group's work with these models, rather than a rigorous quantitative benchmark. Because we have not formally compared the relative data efficiencies of PaiNN and AIMNet2, and are unaware of published literature that provides this specific comparison, we agree that the statement lacks the necessary supporting evidence. Consequently, we have removed this sentence from the revised manuscript:

Designed for generalizability, AIMNet2 natively supports systems with different charge states and spin multiplicities. By explicitly accounting for these varying electronic states and long-range interactions, the model is well-suited for a wide range of chemically complex systems.

L126 P5

There are no citations in the Delta learning section. At least this statement "When the two levels of theory are correlated, this delta-learning approach can substantially reduce model errors." could do with a citation.

We agree with the referee that this statement warrants a citation. We have added a reference to the work by Ramakrishnan et al. (2015), which demonstrated that because the baseline method already captures the underlying physical chemistry, the remaining energy difference is much easier for a machine learning model to predict than learning the target value from scratch. Furthermore, to highlight earlier applications of Δ-learning specifically for molecular dynamics and atomic forces, we have also cited Bogojeski et al. (2020). The revised manuscript now reads:

In this framework, molecular dynamics (MD) simulations are performed at the lower level of theory but are corrected to approximate the high level of theory (Bogojeski et al., 2020). When the two levels of theory are correlated, this delta-learning approach can substantially reduce model errors (Ramakrishnan et al., 2015).

L131 P5

Delta learning section: "The main drawback is that the overall efficiency is fundamentally bounded by the cost of the lower-level baseline." The efficiency of what?

We thank the referee for pointing out this ambiguity. By "efficiency," we were referring to the overall computational cost of the simulation. While the evaluation of the machine learning model is typically fast, a Δ-learning approach also requires evaluating the baseline method (e.g., GFN1-xTB) at every time step. Consequently, the total simulation time is fundamentally bottlenecked by this baseline calculation. This limitation becomes especially pronounced for larger systems, as the computational cost of GFN1-xTB scales with the number of (valence) electrons, whereas the machine learning models scale with the number of atoms. We have

updated the manuscript to clarify this distinction:

The main drawback is that while the evaluation of the machine learning model is typically fast, the overall simulation speed is fundamentally limited by the computational cost of the lower-level baseline.

L134 P5

Data set generation: "Gradient calculations" - What are gradient calculations?

By "gradient calculations," we refer to the computation of the gradients of the potential energy with respect to the nuclear coordinates, which correspond to the negative atomic forces. These force evaluations were executed using the `engrad` keyword in ORCA, as well as the TBLite calculator within the Atomic Simulation Environment (ASE). We have updated the manuscript to explicitly clarify this terminology for a broader audience:

Atomic forces of the selected structures were obtained through potential energy gradient calculations with respect to the nuclear coordinates using GFN1-xTB (TBLite, version 0.2.1) and ωB97X-3c (ORCA, version 6.0.1) (Neese, 2012).

L155 P6

We note that while ωB97X-3c is used in this study, atomic forces can be calculated using any quantum chemistry method on the GFN1-xTB structures to obtain a training set at that level of theory.

L164 P6

Hyperparameter tuning: "we assigned a great weight to the force loss" - Presumably "great" should read "greater" or simply "larger".

We agree that "great weight" is an awkward phrasing and have changed it to "greater weight" as suggested:

..., we assigned a greater weight to the force loss.

L240 P9

Hyperparameter tuning: "This makes sense, as the product of the batch size and batches per epoch determines the total number of samples seen in one epoch." - I am slightly confused. Isn't the definition of an epoch that it is a full run through the data? Once the batch size is determined, the number of batches is set and then simply determines how often the weights are updated in an epoch.

We completely agree that the standard textbook definition of an epoch is exactly one full pass

through the training data. However, the AIMNet2 training framework (built on PyTorch Ignite) operationally defines an "epoch" as a fixed, user-defined number of training steps, detaching it from the actual dataset size.

In our specific setup, our dataset contains 20,000 molecules. With a batch size of 16, a traditional epoch would consist of 1,250 batches. Because the batches per epoch hyperparameter was set to 4,000, the model simply processes 64,000 samples (looping through the dataset multiple times) between each validation step.

We have updated the manuscript to state the operational definition and clarify the exact number of samples processed:

For PaiNN, the batch size and number of features were identified as the most important hyperparameters, with a smaller batch size and a higher number of features correlating with improved performance. In contrast, for AIMNet2, the "batches per epoch" was the dominant hyperparameter. This difference stems from how each training framework defines an epoch. In PaiNN, an epoch follows the standard definition of a single, full pass through the training data. Thus, the training set size and batch size strictly determine the batches per epoch. However, the AIMNet2 framework decouples the definition of an epoch from the dataset size, operationally defining it as a user-specified number of steps before each validation check. Therefore, in AIMNet2, the product of the batch size and the batches per epoch determines the total number of samples processed per operational epoch. When this product is smaller than the training set size, a validation step is triggered before the model has seen all training samples. When the product exceeds the dataset size (e.g., using 1,000 batches of size 16 means the AIMNet2 model processes 16,000 samples per validation cycle for our 2,000-molecule dataset), the model sees data multiple times per epoch, which further aids convergence.

L247 P9

Hyperparameter tuning: "It is important to note that we did not necessarily identify the optimal hyperparameter combination for our systems." — Did you consider using hyperparameter tuning with sample efficient Bayesian optimization as shown e.g., by Stuke et al , Mach. Learn.: Sci. Technol. 2, 035022 (2021)?

We thank the referee for bringing this relevant study to our attention. We had not previously considered sample-efficient Bayesian optimization for hyperparameter tuning. Given the low test errors obtained in the present study, our random grid-search approach proved sufficient for these relatively simple collision systems. However, we completely agree that a more rigorous optimization strategy will be valuable in the future. As we move to more complex simulations, such as simulating multiple particle-forming precursors within a simulation box, we plan to explore this Bayesian optimization.

We revised the manuscript as follows:

It is important to note that we did not necessarily identify the optimal hyperparameter combination for our systems. For instance, while our 100-epoch tuning procedure offers a reasonable indication of training behavior, some hyperparameters might converge slower

but dominate during longer training. Identifying the best training settings would require a systematic search over a broader range of values. While automated techniques such as Bayesian optimization (Stuke et al., 2021) could be explored for more complex systems in future work, the chosen hyperparameters provide sufficiently low test errors for the collision systems studied here, as discussed in the following subsection.

L264 P10

NN training: The accuracies reported in Table 3 are impressive. Already the PaiNN and AIMNet2 models achieve very accurate forces of only a few meV per Angstrom. The Delta-PaiNN model is then even better. What accuracy is actually required for the collision calculations?

We thank the referee for raising this highly relevant question.

As demonstrated in this study and by Knattrup et al. (2025), the long-range capture kinetics for these acid–base systems are not heavily dependent on atomistic specifics and can often be well approximated using simple electrostatic and dispersion potentials. However, our objective here is for the machine learning interatomic potential to accurately describe the entire cluster formation process: the initial approach, the collision, and the subsequent cluster formation through stabilizing proton transfers.

Capturing these events requires stable molecular dynamics trajectories. In the machine learning potential community, a mean absolute force error of 1 kcal/mol/Å is frequently used as a benchmark target for accuracy (e.g., Zaverkin and Kästner, 2021). However, as pointed out by Fu et al. (2022), mean absolute force errors only tell part of the story. The stability of molecular dynamics simulations is sensitive to outliers. Even if the mean errors are low, a single anomalous force prediction in a high-energy region can drive the dynamics off course, potentially leading to catastrophic failures such as unphysical fragmentation.

Because collision trajectories inherently explore these high-energy, non-equilibrium states, minimizing maximum force errors is critical. This is exactly why our study advocates moving beyond global scalar metrics (like mean absolute errors) and instead performing targeted validation against the physical properties of interest, such as spatially resolved errors (e.g., as a function of center-of-mass distance), potentials of mean force, and collision rate coefficients.

Finally, from a computational chemistry perspective, the standard target for thermodynamic properties is "chemical accuracy" (errors < 1 kcal/mol). It is important to distinguish between the error of the method compared to physical reality (the inherent error of the underlying level of theory) and the reproduction error of the machine learning model (which we report in Table 3). Because the total simulation error compounds both of these sources, the machine learning reproduction error must be negligible compared to the inherent error of the reference data. Therefore, achieving energy reproduction errors that fall well below 1 kcal/mol is necessary to ensure the physical outcomes are dictated by the electronic structure method rather than artifacts of the machine learning fit.

Line 265: "This error results from a mismatch between the training labels, which

include long-range stabilization, and the model's short-ranged (< 10 Å) representation. At these distances, the reference energies are significantly lowered by electrostatic interactions. However, due to the 10 Å cutoff, PaiNN interprets the collision partners as two non-interacting, free species. Consequently, during training, the model is forced to attribute the substantial stabilization energy of the interacting pair to the local atomic environments of the isolated monomers. In essence, the model erroneously learns that these structures, separated by more than the cutoff yet still interacting, are representative of the free molecular state, resulting in a fundamentally distorted PES." The analysis sounds interesting, although I would have expected it in the discussion section, but I got a little lost. Figure 2 reports the errors that the authors analyse, but where do we see a mismatch in training labels? And that PaiNN interprets the collision partners as two non-interacting, free species? I feel the reader is given insufficient information to follow the argument.

The referee makes a fair point. The concept we intended to describe is indeed nuanced, and our original phrasing was too abstract and lacked the necessary physical context to easily follow the argument.

The "mismatch" we intended to describe is between the potential energy of the interacting system in the training data and the physical limitations of the model's spatial cutoff. When the two collision partners are, for example, 12 Å apart, the electronic structure calculations in the training data already show a significantly lowered potential energy compared to the isolated monomers due to long-range ion–dipole interactions. These interactions also induce slight geometric distortions in the approaching monomers. To help visualize this, we have provided an illustrative potential of mean force in Fig. R1, showing how the system's energy has already dropped significantly at distances well beyond the 10 Å cutoff.

Because PaiNN has a strict 10 Å spatial cutoff, it evaluates this 12 Å system as two completely isolated, non-interacting monomers. During training, the model must map the significantly lowered potential energy (due to the long-range attractive interactions) to these effectively isolated geometries. Consequently, PaiNN erroneously learns that the slight internal geometric distortion is the cause of the lowered potential energy. It maps the interaction energy onto the isolated monomers, treating their distorted structures as new, highly stable global minima for the free monomers.

We have completely rewritten the paragraph in the manuscript to remove the confusing phrasing and explicitly explain the mechanism of this failure:

In contrast, the ionic H_2SO_4–HSO_4^- system illustrates the inherent limitations of applying a purely local representation to systems with strong long-range interactions. Although the electronic energy MAE for PaiNN appears relatively low, the distance-resolved error plot (Fig. 2c) reveals significant deviations at separations $r > 10$ Å. At these large distances, the training data already capture substantial stabilization energy due to long-range ion–dipole interactions, which also induce slight geometric distortions in the approaching collision partners. However, because PaiNN employs a strict 10 Å spatial cutoff, it evaluates the system as two completely isolated, non-interacting collision partners. During training, the model must map the significantly lowered potential energy of the interacting system to these

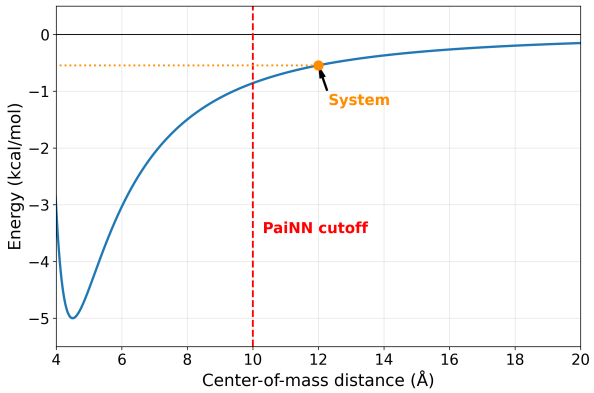

Figure R1: Illustrative potential of mean force highlighting the spatial cutoff limitation. At a center-of-mass distance of 12 Å (orange dot), the interacting system already experiences a lowered potential energy compared to the isolated monomers (Energy = 0 kcal/mol). However, because this distance exceeds PaiNN's 10 Å local atomic environment cutoff (red line), the model evaluates the system as two completely isolated, non-interacting monomers.

isolated, slightly distorted molecular structures. Consequently, PaiNN erroneously learns to associate the long-range electrostatic stabilization entirely with these slight internal structural changes. In essence, the model is forced to view this distorted geometry as the lowest energy conformer of the isolated molecule, creating an artificial global minimum that fundamentally distorts the PES.

L291 P11

Potential of mean force: ***"The potential of mean force (PMF) along the center-of-mass distance represents the effective free energy averaged over all collision orientations accessed during the simulations, showing how the system's stability changes as the collision partners approach. The well depth and shape provide information on the binding strength, while the shoulder towards larger distances reflects the strength of long-range interactions.´´ — Shouldn't one show a PMF as an example or a sketch of a typical PMF so that the reader can follow the statements here? Otherwise it is hard to imagine what the well depth and the shoulder refer to. (I see now that I have read on that PMFs are shown a little later in the section; maybe refer to them here already.)***

We thank the referee for this helpful suggestion. We agree that pointing the reader to a visual example earlier in the text makes the physical interpretation of the potential of mean force features much clearer. We have adjusted the manuscript to explicitly reference the potentials of mean force shown in Fig. 3 when introducing the concepts of the well depth and the shoulder:

The potential of mean force (PMF) along the center-of-mass distance represents the effective free energy averaged over all collision orientations accessed during the simulations, showing how the system's stability changes as the collision partners approach (see e.g., Fig. 3). The

well depth and shape provide information on the binding strength, while the shoulder towards larger distances reflects the strength of long-range interactions.

L303 P12

Table 5: How accurate would the collision rates need to be to be useful in downstream modelling? Presumably the small difference between GFN1-xTB and ωB97X-3c for the neutral dimers is not noticeable, but how about the difference of ∼ 1 for H_2SO_4–H_2SO_4? What I am looking for is some form of contextualisation.

We thank the referee for this question. It is indeed important to contextualize the required accuracy for collision rate coefficients in downstream applications.

To our knowledge, there is currently no community-wide standard for collision rate accuracy analogous to the "chemical accuracy" threshold (1 kcal/mol) used for quantum chemistry energies. Collision rate coefficients are primarily used in cluster distribution dynamics models (such as the Atmospheric Cluster Dynamics Code). In these models, the particle formation rate scales linearly with the collision rate coefficient.

However, formation rates also depend on evaporation rate coefficients, which are calculated from quantum chemically derived binding free energies. Because evaporation rates depend exponentially on these binding energies, errors in the underlying thermochemistry dominate the overall uncertainty. For instance, approaching chemical accuracy (an error of 1 kcal/mol) in binding free energies still introduces a factor of roughly 5 uncertainty in the evaporation rate at room temperature.

A reasonable overarching goal for cluster dynamics simulations is to maintain total particle formation rate errors below one order of magnitude. Allowing for a factor of 5 uncertainty from the evaporation rates, a conservative acceptable error margin for the collision rate coefficients would be a factor of 1.5. Keeping in mind additional sources of error (e.g., improperly defined sources/losses and finite simulation sizes), this factor of 1.5 bounds the compounding error within an order of magnitude.

Evaluated against this threshold, the reproduction errors of all models for all systems fall safely within this threshold, except when the PaiNN model is applied to the charged H_2SO_4–HSO_4^- system, where the error approaches 50% (a factor of ∼2).

We have updated the corresponding paragraph in the manuscript to provide this context:

To contextualize these results, it is important to note that the accuracy of particle formation rates in cluster distribution dynamics simulations depends on both the collision and evaporation rate coefficients. Because evaporation rates depend exponentially on binding free energies, these errors typically outweigh errors in collision rate coefficients. An error of just 1 kcal mol^{-1} in binding free energies introduces a factor of ∼5 uncertainty in the evaporation rate. As such, we consider an error of a factor of 1.5 in the collision rate coefficients acceptable. In the worst-case scenario where collision and evaporation rate coefficient errors compound in the same direction, a factor of 1.5 collision rate coefficient error would still result in an overall uncertainty in the particle formation rates of less than an order of magnitude.

Evaluated against this threshold, PaiNN yields notably lower rate coefficients for the charged H_2SO_4–HSO_4^- system. For both GFN1-xTB and ωB97X-3c training data, the model underestimates the rates by nearly 50% (roughly a factor of 2) relative to the GFN1-xTB reference. As discussed in Sec. 3.5, this substantial deviation stems from the model's inability to detect collisions beyond its 10 Å cutoff, effectively neglecting the significant contribution of the long-range tail.

L390 P18

References

Fu, X., Wu, Z., Wang, W., Xie, T., Keten, S., Gomez-Bombarelli, R., and Jaakkola, T.: Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations, arXiv preprint arXiv:2210.07237, https://doi.org/10.48550/arXiv.2210.07237, 2022.

Knattrup, Y., Neefjes, I., Kubečka, J., and Elm, J.: Growth of atmospheric freshly nucleated particles: a semi-empirical molecular dynamics study, Aerosol Research, 3, 237–251, https://doi.org/10.5194/ar-3-237-2025, 2025.

Zaverkin, V. and Kästner, J.: Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design, Machine Learning: Science and Technology, 2, 035 009, https://doi.org/10.1088/2632-2153/abe294, 2021.